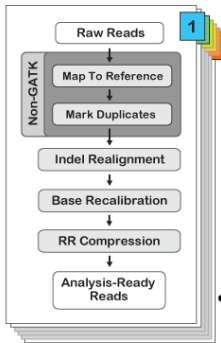# Variants

Martin Morgan (mtmorgan@fhcrc.org)
Fred Hutchinson Cancer Research Center
Seattle, WA

4 February 2014
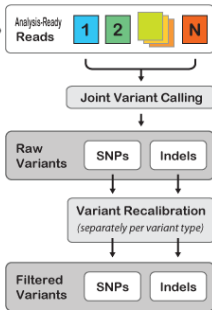
Work flows

1. Experimental design – tumor / normal pairs; cell lines; . . .
2. Sequencing – DNA or Exome
3. Alignment & other pre-processing steps
4. Variant discovery & preliminary analysis
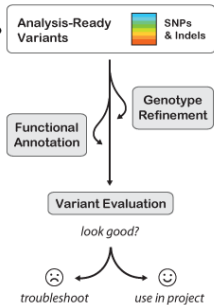5. **Variant evaluation, annotation, biological and experimental context**

Data Pre-processing   >>   Variant Discovery   >>   Preliminary Analyses

Non-GATK

Raw Reads

Map To Reference

Mark Duplicates

Indel Realignment

Base Recalibration

RR Compression

Analysis-Ready Reads

Analysis-Ready Reads   1   2   N

Joint Variant Calling

Raw Variants   SNPs   Indels

Variant Recalibration
(separately per variant type)

Filtered Variants   SNPs   Indels

Analysis-Ready Variants   SNPs & Indels

Functional Annotation

Genotype Refinement

Variant Evaluation

look good?

troubleshoot   use in project

Variant Call Format (VCF) files

- Specification
- Header documenting file content
- CHROMosome, POSition, IDentifier of each variant
- REFerence and ALTernate allele sequence.
- INFOrmation on variants
- FORMAT of sample genotype information, followed by each genotype

VCF content: location

```
#CHROM POS       ID        REF     ALT    QUAL FILTER ...
20     14370     rs6054257 G       A      29   PASS   ...
20     17330     .         T       A      3    q10    ...
20     1110696   rs6040355 A       G,T    67   PASS   ...
20     1230237   .         T       .      47   PASS   ...
20     1234567   microsat1 GTC     G,GTCT 50   PASS   ...
```

Lines: good SNP, poor quality SNP, multipe variants, called monomorphic, indel

VCF content: variant INFO

```
#CHROM POS      ... INFO                                  ...
20     14370    ... NS=3;DP=14;AF=0.5;DB;H2                ...
20     17330    ... NS=3;DP=11;AF=0.017                    ...
20     1110696  ... NS=2;DP=10;AF=0.333,0.667;AA=T;DB      ...
20     1230237  ... NS=3;DP=13;AA=T                        ...
20     1234567  ... NS=3;DP=9;AA=G                         ...
```

Information supporting the SNP: NS, # samples with data; DP, total depth; AF, ancestral frequency; DB, dbSNP membership; H2, HapMap 2 membership.

VCF content: Genotype FORMAT and samples

```
... POS     ... FORMAT      NA00001        NA00002
... 14370   ... GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51
... 17330   ... GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3
... 1110696 ... GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
... 1230237 ... GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51
... 1234567 ... GT:GQ:DP    0/1:35:4       0/2:17:2
```

Genotype inforamation in three samples. FORMAT specifies the
order and type of information: GT, Genotype, '|' phased, vs. '/'
unphased; GQ, quality; DP, read depth; HQ, haplotype quality.

VCF Header

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f
##phasing=partial
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency"
...
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data
...
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Qual
```

VCF files

- ▶ Very complicated data.
- ▶ Content of INFO, FORMAT fields very flexible, depends entirely on up-stream processing.
- ▶ Often interested in only part of the file – specific genomic ranges, INFO or FORMAT fields, samples.